Journal of Computer and Knowledge Engineering, Vol. 8, No. 1, 2025. (35-42)

Ferdowsi University of Mashhad

Journal of Computer and Knowledge Engineering

https://cke.um.ac.ir

Information and Communication Technology Association of Iran

PerGOLD: Identification of offensive language in Persian tweets: leveraging crowdsourcing*

Fatemeh Jafarinejad[1], Marziea Rahimi[2], Maryam Khodabakhsh[3], and Seyedehfatemeh Karimi[4]

10.22067/cke.2025.90088.1132

Abstract *It is concerning that the growing popularity of social networks is encouraging violence or inciting offense toward other people. An attempt has been made in the past several years to detect offensive language in social media posts. Nonetheless, the majority of studies focus on recognizing offensive language in English. Moreover, dataset labeling emerges as a crucial and fundamental step for training high-quality models, considering the increasing use of artificial intelligence and machine learning tools. Utilizing crowdsourcing platforms is an efficient and optimal method that can be used for data labeling. This approach uses human resources who are sufficiently knowledgeable about the topic to label the data. In this paper, we introduce PerGOLD, a new Persian General Offensive Language Dataset, in which we use an event-based data collection methodology to detect offensive language in Persian Twitter. To access labeled training data, we build a crowdsourcing platform to benefit from human input. We labeled 13,716 tweets, and according to the obtained results, 34% of them were labeled as offensive language. Finally, we evaluated the efficiency of these data by applying some classic machine learning models (LR, SVM) and transformer-based language models (RoBERTa, ParsBERT). The obtained F1-score of the best model (ParsBERT) was 85.4%.*

Key Words: Offensive Language, Labeling, Crowdsourcing, Natural Language Processing.

1. INTRODUCTION

Social networks are widely used by individuals these days to share their experiences, activities, and opinions with others, freely and anonymously. While freedom of expression is a fundamental right for all individuals, the expression of Offensive Language (OL) constitutes an abuse of this freedom [1][2]. The social platforms are responsible for these abusive cases, which will confront communities with various social issues. The speed of information production in social networks and the web, as well as the relative freedom and anonymity of cyberspace, are some issues making manual diagnosis difficult or impractical. Qian et al. defined OL as language that includes toxic, hateful, abusive, violent, and bullying characteristics [3].

In the automatic detection of OL, data preparation is one of the fundamental stages. Influencing the quality of the knowledge extraction process and improving the accuracy of Machine Learning (ML) models [4]. Thus, data labeling is a fundamental and significant issue for researchers in the field of ML and data mining since labeled datasets are utilized as inputs in algorithms and ML models [5]. People are more efficient and perform better than computers in many activities, including recognizing concepts in texts and images, categorizing documents, translating natural language, and evaluating the value of items [6].

Nowadays, various studies have been conducted to introduce datasets for identifying offensive language [7]. Most studies collected their datasets from social media and then manually labeled them based on task requirements. Twitter is also the most popular platform used for dataset collection [8]. The labeling process has been done with the help of experts [9], native speakers [10], volunteers [11], or crowdsourcing from users [12]. Small dataset sizes, lack of offensive content ratios, and lack of label definitions and agreements among taggers [8] are among the issues addressed for OL. These issues are stronger in low-resource languages, e.g. Persian.

Crowdsourcing, as one of the most effective and common approaches to data labeling, employ humans to perform tasks in exchange for rewards, honor, or entertainment purposes [13][14]. Crowdsourcing-based platforms send tasks to humans and then collect the results.

* Manuscript received 2024 October 7, Revised 2024 November 3, accepted 2025 February 22.

[1] Assistant Professor, Department of the Computer Engineering, Shahrood University of Technology (SUT), Shahrood, Semnan, Iran.

[2] Assistant Professor, Department of the Computer Engineering, Shahrood University of Technology (SUT), Shahrood, Semnan, Iran.

[3] Corresponding author. Assistant Professor, Department of the Computer Engineering, Shahrood University of Technology (SUT), Shahrood, Semnan, Iran, **Email:** m_khodabakhsh@shahroodut.ac.ir

[4] Master Graduate, Department of Computer Engineering Ferdowsi University of Mashhad, Mashhad, Iran.

The crowdsourcing systems consist of 4 main components[14]: 1) Requester, 2) Worker, 3) Task, and 4) Platform. Requesters are recognized as a group of individuals, publishers, or seekers. The requester is the system administrator who defines the labeling task and conditions, and we refer to them as Admin. Workers are recognized as individuals with the knowledge and skills to solve problems or participate in projects for outsourced tasks. The workers are the users who are responsible for labeling the data, and we refer to them as Tagger. The task describes an object that is outsourced or a group of instructions. Determining the appropriate labels for the available data is defined as a task in our work. The platform serves as an interface between the requester and the workers. The platform provides efficient actions for organizing and managing the entire crowdsourcing process and may handle some tasks related to the requesters. In this study, a crowdsourcing platform has been implemented in the form of an online website[5].

The main goal of our work is to identify OL and label the data. We address this by creating a crowdsourcing system for labeling the data extracted from Twitter. Our contributions in this article can be summarized as follows:

- Developing an OL dataset for the Persian language as a low-resource language.
- Using an event-based approach for tweet collection, which leads to a broader variety of offensive language samples and a higher proportion of offensive tweets.
- Developing a crowdsourcing platform for tweet labeling.
- Conducting extensive experiments using classic classifiers and transformer language models to demonstrate the validity of the data for offensive language detection.

We conduct experiments by training some classical and state-of-the-art ML models to demonstrate the quality of the labeling achieved through crowdsourcing. Results demonstrate the effectiveness of our proposed framework in this task. The rest of the paper is organized as follows: Section 2 describes related works conducted in the field of data labeling in OL. Then, in Section 3, we provide detailed explanations about the developed crowdsourcing-based system. Subsequently, in order to evaluate the system, we employed it in an application related to examining OL on the Twitter social network, which is discussed in Section 4. We introduce the dataset, the labeling process, the models used, and evaluate the results obtained from them. Lastly, conclusion and future work are discussed in Section 5.

2. RELATED WORK

OL is becoming more widespread along with the growth of online content. Social media platforms play a significant role in the production of OL [2]. Problematic user-generated content can include aggressive, threatening, poisonous, misogynistic, abusive, insulting language, and OL [15]. Many online platforms like Facebook, YouTube, and Twitter consider OL harmful and have approaches in place to remove content that promotes hate from their platforms [2]. Given that OL can lead to social problems [16], studying the detection of OL is crucial. However, most studies focus on detecting OL based on the English language [17]. Therefore, the existence of studies that focus on OL detection in low-resource languages is essential.

Detecting OL comes with challenges. One challenge is the disagreement in defining OL [2]. Legal frameworks, social science research, and social media companies have definitions for OL based on their own objectives [18]. However, there is no universally accepted global definition of OL [17]. This causes a piece of content to be considered OL by some, while others do not view it as such [2]. A number of factors have been linked to people's perceptions of OL, including gender, ethnicity, racial attitudes, the value placed on freedom of speech, context, target, empathy, ways of knowing, implicitness of the OL, and relationality [15].

As OL increases in social networks, it becomes critical to train appropriate models for OL identification. Training such models requires labeled datasets. However, labeled datasets are generally scarce, particularly for low-resource languages [19]. In Table 1, we review some OL datasets in terms of language, volume, data collection platform, and the models used for dataset evaluation. These datasets have been created for various languages such as Arabic [20][21], Greek [20], Turkish [20], Urdu [22], and Persian [7][19][23][24], each having different sizes based on the data collection and labeling methods. Additionally, posts and user comments from different platforms such as Twitter [19][21][22][23][24] and Instagram [7][23] have been utilized to create these datasets. Various studies have employed different classification methods to evaluate the existing datasets. Some studies have evaluated the performance of their datasets using basic text classification methods, while others have utilized neural networks or a combination of different classifiers.

Among the Persian datasets in Table 1, our dataset, consisting of 13,716 tweets, ranks second in terms of volume after the OPSD dataset [23], which contains 17,000 tweets. Therefore, PerGOLD can be considered one of the biggest datasets of offensive tweets. On the other hand, we have used event-based tweet collection, which is the most significant difference between our dataset and similar Persian datasets. The event-based tweet collection enabled us to gather a wider range of offensive comments, resulting in a higher proportion compared to non-offensive tweets. Offensive language is often triggered by social and political events [25]. As a result, offensive comments related to these events are more frequent and varied. Collecting tweets associated with such events creates a dataset that includes a broader range of offensive sentences with a more natural distribution.

Ali et al. [22] are engaged in collecting and labeling Urdu tweets with the aim of sentiment analysis. They extracted a collection of tweets over a period of 6 months by searching for appropriate hashtags. In order to obtain a suitable dataset, they also worked on correcting the tweets

[5] http://intelligeco.ir

in terms of grammar. After collecting the data, they sought the help of experts to label the data and determine the type of tweet context (national security, religious, and ethnic differences). In the end, they collected a dataset with 16,000 records and used SVM[6] and MNB[7] models to evaluate the performance of their dataset.

Çöltekin [26] introduced the first labeled dataset for OL in the Turkish language. Their final dataset consists of 36,232 tweets collected over a period of 18 months. They used a hierarchical structure, presented in [27], for labeling the tweets, initially dividing them into two categories, the OL category or the non-OL category. Furthermore, they determined whether the offensive tweets were not targeted or targeted. Finally, for targeted tweets, they specified the type of target, whether it was an individual, a group, or others. They requested annotators to assign one or more labels to each tweet based on this hierarchical structure. After labeling the data, they employed the SVM model to examine the performance of the labeling process.

Khodabakhsh et al. [7] conducted the collection and labeling of a dataset extracted from Persian comments on Instagram. They utilized user-based and news agencies-based approaches to gather the data. Then, they employed three annotators to label the data into three categories, offensive, non-offensive, and Advertisement. For the offensive data, they further specified the type of insult, including curse, insult, sexist, origin, racist, national, religion, political, and sexual. In the end, they collected a labeled dataset consisting of 28,164 records and evaluated it using BNB[8], GNB[9], and LR[10] models.

3. Crowdsourcing Platform

Data and the quality of its labels are the most important aspects of this study. Therefore, we have designed a crowdsourcing system to manage the labeling process. The architecture of this system can be seen in Figure 1. In this system, there are three entities, including Task Owner, tagger, and user. A task is an outsourced object with a set of instructions [14]. In our system, data labeling is defined as a task that includes definitions and guidelines for high-quality data labeling, based on the intended purpose. The task owner, also known as the requester, publishes tasks to find solutions to specific issues [14]. In our system, they can perform functions such as defining tags for projects, setting up tasks, and supervising member performance. The Tagger, or worker, possesses the knowledge and skills needed to complete outsourced tasks [14]. In our system, they can access the labeling section, label consecutive displayed sentences, and review their own performance. Users are a large group who join our system. They can participate as taggers in data labeling tasks if approved by task owner. Additionally, users have the ability to enter a sentence and receive the corresponding label for it, in addition to browsing website pages.

In our system, tasks are designed by task owner. Moreover, task owner uploads data for labeling into the system and assigns taggers to perform specific tasks. To enhance the quality of labeling, task owner has the ability to select taggers. This means that task owner can choose individuals among the users who have the necessary knowledge and skills to perform the desired task as taggers. After defining tasks and taggers, the labeling phase begins. The sequence diagram depicting the process of labeling the data can be seen in Figure 2. First, tagger logs into their dashboard by entering their username and password, and then they enter the labeling page. Next, a sentence is presented to tagger, who can choose one of the predefined labels. Then the system verifies the selected label and stores it in the database. After saving the labeled sentence, a new unlabeled sentence is presented to tagger.

TABLE 1
Comparison of OL datasets

Paper	Language	Number of records	Social media	Models
[28]	Arabic	10,000	Twitter	BERT, AraBERT, DT, RF, GNB, AdaBoost, Perceptron, Gradient Boosting, LR, SVM
[29]	Greek	10,228	Twitter	GRU, LSTM, BERT
[26]	Turkish	35,284	Twitter	SVM
[22]	Urdu	16,000	Twitter	SVM, MNB
[21]	Arabic	11,000	Twitter	SVM, LTSM, CNN + LTSM, GRU, CNN + GRU
PerBOLD [7]	Persian	28,164	Instagram	Bernoulli NB, Gaussian NB, LR
[19]	Persian	7,056	Twitter	ParsBERT, mBERT, XML-R, ChatGPT
[23]	Persian	21,165	Twitter, Instagram	ALBERT-fa, ParsBERT, RoBERTa-fa, XLM-RoBERTa
[24]	Persian	8,013	Twitter	LR, SVM, CNN
PerGOLD	Persian	13,716	Twitter	LR, SVM, RoBERTa

[6] Support Vector Machine
[7] Multinomial Naive Bayes
[8] Bernoulli Naive Bayes
[9] Gaussian Naive Bayes
[10] Logistic Regression

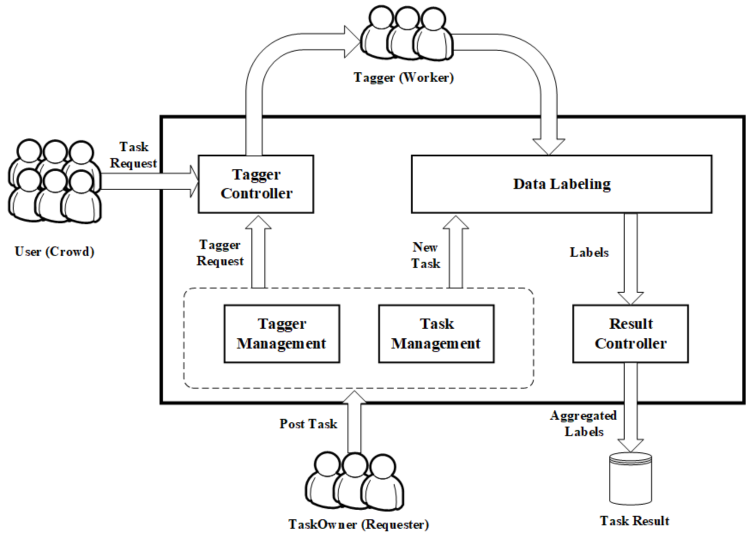


Figure 1. The architecture of the crowdsourcing system for data labeling

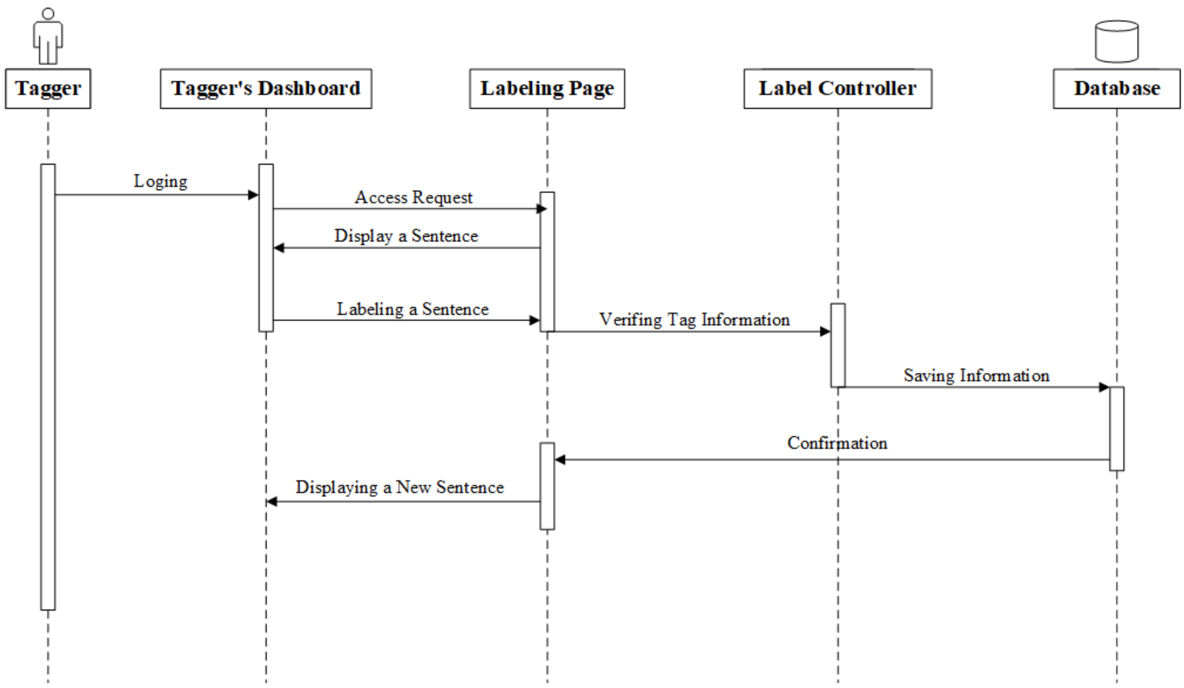


Figure 2. The sequence diagram related to the labeling process

4. RESULTS AND DISCUSSION

4.1. Data Collection and Annotation

Nowadays, millions of users freely express their thoughts and opinions on Twitter, leading to an increase in OL on this social media platform [19]. For this reason, in this study, we focused on collecting a significant dataset of OL based on data retrieved from Twitter. Various strategies have been proposed in studies to collect relevant data from social networks. Most studies use a variety of data collection strategies, including keyword-based, user-based, news agency-based, and event-based [7]. We employed an event-based approach along with keyword selection to gather the data from the Twitter social network.

Additionally, given the importance of studying OL in low-resource languages, we focused on extracting Persian tweets. Procedures used to create the labeled dataset are illustrated in Figure 3. Initially, a dataset containing approximately 14k Persian tweets was extracted and placed in the crowdsourcing system to obtain labels of whether they were OL or not.

To collect the data, we examined various events that had triggered OL among Persian Twitter users. Subsequently, we focused on extracting tweets within a 14-day period following each event, using appropriate keywords related to that event. Furthermore, in order to create a comprehensive dataset, we attempted to consider relevant events for each different type of OL, like ethnic, national, origin and lineage, gender, religious, racial, and physical condition. Then we proceeded with extracting tweets using suitable keywords for each type of OL.

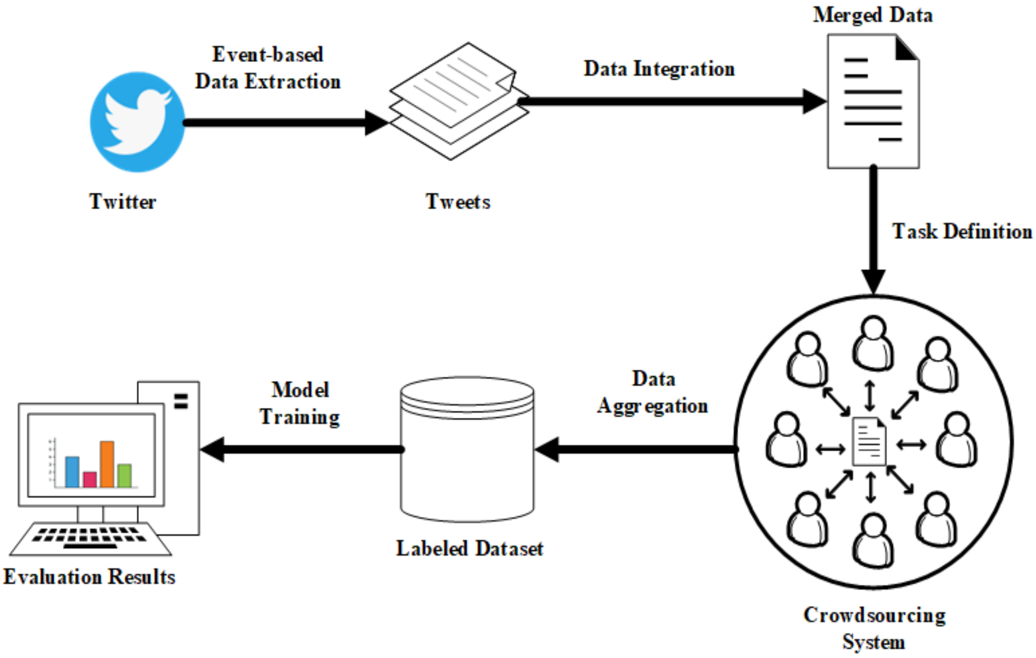


Figure 3. Workflow process for the creation of the labeled dataset

After preparing the tasks in the system, the labeling process of the data was carried out by defining 4 taggers. To ensure consistency in labeling, a labeling guideline was created with the following general rules:

- Plain or promotional texts are categorized as non-OL.

Texts containing insults towards sacred beliefs, provocative sexual/gender discussions, racial insults, or insults towards disabled individuals are classified as OL.

4.2. Statistics and Experiments

In PerGOLD, a total of 13,716 data points from the Twitter dataset received OL/non-OL labels, with 9,103 data points categorized as non-OL and 4,613 as OL. It is worth noting that approximately 34% of the entire collection consists of OL. Moreover, data labeling in this study was performed in a binary manner, meaning that each tweet was assigned only one label of OL or non-OL.

Table 2 provides a comparison of OL and non-OL data in different datasets, highlighting the distribution of these categories. In most of the presented datasets, the ratio of OL to non-OL data is low. The imbalance in the number of data necessitates anomaly detection and further removal of data categories, resulting in data loss. In this research, we attempted to focus on extracting tweets with a higher percentage of OL by finding suitable events and keywords. Ultimately, our dataset contains 34% OL data, indicating the effectiveness of our data collection methodology.

In order to be accompanied, we apply some ML as some baseline text categorization models to demonstrate the coherence and effectiveness of the proposed dataset. For this purpose, we first preprocessed the data by applying preprocessing techniques such as removing numbers, web addresses, emails, monetary units, emojis, and punctuation marks. The data has been preprocessed using the Hazm library. Subsequently, the data was divided into two sets, with 70% for training and 30% for testing. A binary classification task was performed, distinguishing between non-OL and OL. Following that, two classical ML models (LR, SVM) and two large language models (LLMs) (RoBERTa [34], ParsBERT [35]) have been trained. Table 3 displays the parameters that were utilized in the model training process.

TABLE 2

The statistical information related to various OL datasets

Dataset	Language	Number of records	OL	Non-OL
[30]	English	6,000	1,567	4,433
[31]	English	31,962	2,242	29,720
[28]	Arabic	10,000	1,915	8,085
[32]	Danish	3,600	441	3,159
[29]	Greek	10,287	2,911	7,376
[26]	Turkish	35,284	6,847	28,464
[33]	Persian	6,000	1,624	4,376
PerGOLD	Persian	13,716	4,613	9,103

Table 4 shows the performance of applying these models to test data. In terms of accuracy, precision, recall, and F1-score. The results show that when it comes to OL classification on our labeled data, the LR and transformer-based models perform better than the SVM model. In the case of transformer-based models, ParsBERT, which is a monolingual language model with BERT architecture, outperforms RoBERTa as a multilingual architecture. ParsBERT trains the BERT from scratch with a corpus of more than 3.9 million Persian documents.

This result provides valuable information about training

an architecture in a monolingual or multilingual manner: In monolingual NLP tasks, using an architecture and training it from scratch with the development of the corresponding monolingual corpus may be more successful in acquiring and encoding language-specific knowledge than using multilingual models with fine tuning. Based on the results obtained on the dataset, the ParsBERT model exhibits the best performance among the classic and transformer-based models with a F1-score of 85.4%, while the RoBERTa has a performance near a classic model (LR).

TABLE 3
Features of trained models

Model	Learning Technique	Architecture	Parameters	Iterations
LR	Classical ML	Logistic Regression classification	CW = 2 Optimizer = LBFGS	1000
SVM	Classical ML	Support Vector Machine	C = 0.1	10
RoBERTa	Transformer-based Language model	Roberta-base with Adam Optimizer	BatchSize = 32 Learning Rate = 2E-5	3
Hooshvare-1	Transformer-based Language model	BERT, Tuned On Persian Dataset	BatchSize=4 Learning Rate= 2E-6 NumLayers = 6	3
Hooshvare-2	Transformer-based Language model	BERT, Tuned On Persian Dataset	BatchSize=4 Learning Rate= 2E-6 NumLayers = 12	3

TABLE 4
Results of applying some baseline categorization models on the dataset

Model	Accuracy	Precision	Recall	F1- score
LR	0.75	0.73	0.74	0.73
SVM	0.5	0.25	0.5	0.33
RoBERTa	0.71	0.72	0.72	0.71
ParsBERT-1	0.83	0.83	0.82	0.824
ParsBERT-2	**0.85**	**0.86**	**0.85**	**0.854**

5. Conclusion

In this paper, we present the PerGOLD, a dataset of the Tweeter comments in Persian. Utilizing the event-based methodology in data collection, a 14-day period of different events triggering offences was searched by appropriate keywords. Subsequently, we implemented a crowdsourcing platform to label the data in a 2-class classification task. In order to be accompanied, we apply baseline text categorization models (LR, SVM, RoBERTa, ParsBERT) to demonstrate the coherence and effectiveness of the proposed dataset. The experimental results illustrate that in this language-specific task, a monolingual language-specific model (ParsBERT) outperforms other models in acquiring and encoding language-specific knowledge. On the other hand, multilingual models (RoBERTa) perform as a classic model (LR).

6. References

[1] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti. (2021, Jun.). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*. [Online]. 55, pp. 477–523. Available: https://doi.org/10.1007/s10579-020-09502-8

[2] S. MacAvaney, H. R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder. (2019, Aug.). Hate speech detection: Challenges and solutions. *PLoS One*. [Online]. 14(8), p. e0221152. Available: https://doi.org/10.1371/journal.pone.0221152

[3] J. Qian, A. Bethke, Y. Liu, E. Belding, and W. Y. Wang. (2019, Sep.). A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251*. [Online]. Available: https://doi.org/10.48550/arXiv.1909.04251

[4] J. Zhang, V. S. Sheng, T. Li, and X. Wu. (2017, Mar. 22). Improving crowdsourced label quality using noise correction. *IEEE Transactions on Neural Networks and Learning Systems*. [Online]. 29(5), pp. 1675–1688. Available: https://doi.org/10.1109/TNNLS.2017.2670702

[5] C. Li, V. S. Sheng, L. Jiang, and H. Li. (2016, Sep.). Noise filtering to improve data and model quality for crowdsourcing. *Knowledge-Based Systems*. [Online]. 107, pp. 96–103. Available: https://doi.org/10.1016/j.knosys.2016.06.006

[6] W. Tang and M. Lease. (2011, Jul.). Semi-supervised consensus labeling for crowdsourcing. In SIGIR 2011 Workshop on Crowdsourcing for Information

Retrieval (CIR), pp. 1–6. [Online]. Available: https://mattlease.com/papers/tang-cir11.pdf

[7] M. Khodabakhsh, F. Jafarinejad, M. Rahimi, and M. Ghayoomi. (2023). PerBOLD: A Big Dataset of Persian Offensive Language on Instagram Comments. *Tabriz Journal of Electrical Engineering*. [Online]. 53(2), pp. 149-158. Available: https://doi.org/10.22034/tjee.2023.15794

[8] M. S. Jahan and M. Oussalah. (2023, Aug.). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*. [Online]. 546, pp. 126232. Available: https://doi.org/10.1016/j.neucom.2023.126232.

[9] Z. Talat and D. Hovy. (2016, Jun.). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In Proceedings of the NAACL Student Research Workshop, pp. 88–93. [Online]. Available: https://aclanthology.org/N16-2013.pdf

[10] L. Gao and R. Huang. (2017, Oct.). Detecting online hate speech using context aware models. *arXiv preprint arXiv:1710.07395*. [Online]. Available: https://doi.org/10.26615/978-954-452-049-6_036

[11] W. Warner and J. Hirschberg. (2012, Jun.). Detecting hate speech on the world wide web. In Proceedings of the Second Workshop on Language in Social Media, pp. 19–26. [Online]. Available: https://aclanthology.org/W12-2103.pdf

[12] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. (2019, Mar.). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*. [Online]. Available: https://doi.org/10.48550/arXiv.1903.08983

[13] P. Venetis, H. Garcia-Molina, K. Huang, and N. Polyzotis. (2012, Apr.). Max algorithms in crowdsourcing environments. In Proceedings of the 21st International Conference on World Wide Web, pp. 989–998. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/2187836.2187969

[14] S. S. Bhatti, X. Gao, and G. Chen. (2020, Sep.). General Framework, Opportunities and Challenges for Crowdsourcing Techniques: A Comprehensive Survey. *Journal of Systems and Software*. [Online].167, p. 110611. Available: https://doi.org/10.1016/j.jss.2020.110611

[15] Y. Sang and J. Stanton. (2022, Feb.). The origin and value of disagreement among data labelers: A case study of individual differences in hate speech annotation. In International Conference on Information, pp. 425–444. Cham: Springer International Publishing. [Online]. Available: https://doi.org/10.1007/978-3-030-96957-8_36

[16] N. Meedin, M. Caldera, S. Perera, and I. Perera. (2022). A novel annotation scheme to generate hate speech corpus through crowdsourcing and active learning. *International Journal of Advanced Computer Science and Applications*. [Online]. 13(11). Available: https://www.academia.edu

[17] F. Poletto, M. Stranisci, M. Sanguinetti, V. Patti, and C. Bosco. (2017). Hate speech annotation: Analysis of an Italian Twitter corpus. CEUR Workshop Proceedings, 2006. [Online]. Available: https://doi.org/10.4000/books.aaccademia.2448

[18] A. Sellars. (2016). Defining Hate Speech. *Berkman Klein Center Research Publication* 2016-20. [Online]. pp.16-48. Available: https://doi.org/10.2139/ssrn.2882244

[19] Z. Delbari, N. S. Moosavi, and M. T. Pilehvar. (2024). Spanning the Spectrum of Hatred Detection: A Persian Multi-Label Hate Speech Dataset with Annotator Rationales. In Proceedings of the AAAI Conference on Artificial Intelligence, pp. 17889–17897. [Online]. Available: https://doi.org/10.1609/aaai.v38i16.29743

[20] A. Safaya, M. Abdullatif, and D. Yuret. (2020, Jul.). Kuisail at SemEval-2020 Task 12: BERT-CNN for offensive speech identification in social media. *arXiv preprint arXiv:2007.13184*. [Online]. Available: https://doi.org/10.48550/arXiv.2007.13184

[21] A. Al-Hassan and H. Al-Dossari. (2022, Dec.). Detection of hate speech in Arabic tweets using deep learning. *Multimedia Systems*. [Online]. 28(6), pp. 1963–1974. [Online]. Available: https://doi.org/10.1007/s00530-020-00742-w

[22] M. Z. Ali, S. Rauf, K. Javed, and S. Hussain. (2021, Jun.). Improving hate speech detection of Urdu tweets using sentiment analysis. *IEEE Access*.[Online]. 9, pp. 84296–84305. Available: https://doi.org/10.1109/ACCESS.2021.3087827

[23] M. Safayani, A. Sartipi, A. H. Ahmadi, P. Jalali, A. H. Mansouri, M. Bisheh-Niasar, and Z. Pourbahman. (2024, Apr.). OPSD: an Offensive Persian Social Media Dataset and its baseline evaluations. *arXiv preprint arXiv:2404.05540*. [Online]. Available: https://doi.org/10.48550/arXiv.2404.05540

[24] M. K. Sheykhlan, J. Shafi, and S. Kosari. (2023, Oct.). Pars-HAO: Hate speech and offensive language detection on Persian social media using ensemble learning. *Authorea Preprints*. [Online]. Available: https://doi.org/10.36227/techrxiv.24106617.v2

[25] W. T. Tufa, I. Markov, and P. Vossen. (2024, Jun.). Grounding Toxicity in Real-World Events Across Languages. In International Conference on Applications of Natural Language to Information Systems, pp. 197–210. Cham: Springer Nature Switzerland. [Online]. Available: https://doi.org/10.1007/978-3-031-70239-6_14

[26] Ç. Çöltekin. (2020, May.). A corpus of Turkish offensive language on social media. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 6174–6184. [Online]. Available: https://aclanthology.org/2020.lrec-1.758/

[27] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. (2019). Predicting the Type and Target of Offensive Posts in Social Media. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 1415–1420. [Online]. Available: https://doi.org/10.48550/arXiv.1902.09666

[28] H. Mubarak, A. Rashed, K. Darwish, Y. Samih, and

A. Abdelali. (2020, Apr.). Arabic offensive language on Twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*. [Online]. Available: https://doi.org/10.48550/arXiv.2004.02192

[29] Z. Pitenis, M. Zampieri, and T. Ranasinghe. (2020, Mar.). Offensive anguage identification in Greek. *arXiv preprint arXiv:2003.07459*. [Online]. Available: https://arxiv.org/abs/2003.07459

[30] L. Quijano-Sanchez, J. C. P. Kohatsu, F. Liberatore, and M. Camacho-Collados. (2019). HaterNet: A system for detecting and analyzing hate speech in Twitter. Zenodo. [Online]. Available: https://doi.org/10.5281/zenodo.2592149

[31] P. K. Roy, A. K. Tripathy, T. K. Das, and X. Gao. (2020). A Framework for Hate Speech Detection Using Deep Convolutional Neural Network. IEEE Access, pp. 204951–204962. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.3037073

[32] G. I. Sigurbergsson and L. Derczynski. (2019, Aug.). Offensive language and hate speech detection for Danish. *arXiv preprint arXiv:1908.04531*. [Online]. Available: https://doi.org/10.48550/arXiv.1908.04531

[33] M. Mozafari, S. Member, and R. Farahbakhsh. (2022). Cross-Lingual Few-Shot Hate Speech and Offensive Language Detection Using Meta Learning. *IEEE Access*. [Online]. 10, pp. 14880–14896. Available: https://doi.org/10.1109/ACCESS.2022.3147588

[34] Y. Liu et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*. [Online]. 364(1). Available: http://arxiv.org/abs/1907.11692

[35] M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri. (2021, Dec.). ParsBERT: Transformer-based model for Persian language understanding. *Neural Processing Letters.* [Online]. 53, pp. 3831–3847. Available: https://doi.org/10.1007/s11063-021-10528-4